北京理工大學

课程作业(论文)

YOLO 系列目标检测模型综述

Overview of YOLO series object detection models

小　　组：B11

成　　员：赵思齐，刘京源，蔡康乾，杨昕璋，王涵驰

课程名称：计算机科学与技术前沿

主讲教师：张宇霞，韩锐

2024 年 11 月 19 日

YOLO 系列目标检测模型综述

摘要

随着计算机视觉技术的快速发展，目标检测在各个领域中发挥着越来越重要的作用。YOLO（You Only Look Once）系列模型作为一种高效的实时目标检测方法，因其优越的速度和准确性而广受关注。本文综述了 YOLO 系列模型的发展历程，包括 YOLOv1 到 YOLOv10 的主要架构和创新点。我们分析了每个版本在网络设计、特征提取、模型优化以及量化技术等方面的改进，探讨了它们在不同数据集上的性能表现。此外，本文还讨论了 YOLO 系列模型在实际应用中的优势与局限性，并展望了未来的研究方向。通过对 YOLO 系列的系统性回顾，旨在为研究人员和工程师提供一个全面的参考，以推动目标检测技术的进一步发展。

关键词：YOLO；目标检测；深度学习；计算机视觉；模型优化

abstract

With the rapid development of computer vision technology, object detection plays an increasingly important role in various fields. The YOLO (You Only Look Once) series models have gained significant attention as an efficient real-time object detection method due to their superior speed and accuracy. This paper reviews the development of the YOLO series, covering the key architectures and innovations from YOLOv1 to YOLOv7. We analyze the improvements in network design, feature extraction, model optimization, and quantization techniques in each version, and evaluate their performance on different datasets. Additionally, this paper discusses the advantages and limitations of the YOLO series models in practical applications and explores future research directions. Through a systematic review of the YOLO series, we aim to provide a comprehensive reference for researchers and engineers to further advance object detection technology.

Keywords：YOLO; object detection; deep learning; computer vision; model optimization

目录

第 1 章　研究背景

1.1 引言

目标检测是计算机视觉中的重要任务，广泛应用于自动驾驶、安防监控和医疗影像等领域。[4]其挑战在于同时完成目标定位和分类，尤其在复杂场景中要求模型兼顾精度与效率。基于深度学习的目标检测方法分为两阶段（如 Faster R-CNN）和单阶段（如 YOLO）两类，其中 YOLO 系列以高效性和实时性脱颖而出，成为单阶段检测的代表。

自 2016 年提出以来，YOLO 系列不断演进，通过多尺度检测、特征融合和优化策略，提升了模型在复杂任务中的表现。本文综述了 YOLO 系列模型的技术演化及关键创新，分析其在实际应用中的优势与局限，并探讨未来发展方向。

1.2 目标检测的历史与重要性

目标检测是计算机视觉领域中的核心问题之一，旨在从图像或视频中检测并标注感兴趣的目标，同时预测其类别和位置。在计算机视觉中，目标检测一直是个很有吸引力的领域。它的目标是在图像或视频中找到感兴趣的目标，并且已经应用于许多现实世界的应用中。[1]近年来，随着更深层、更复杂的网络架构的提出，YOLO 算法的特征提取能力不断加强，YOLO 算法经过长时间的发展，已迭代到v9版本。图 1-1 展示了YOLO算法的时间发展线。

早期目标检测方法主要依赖手工设计特征和传统机器学习算法（如HOG + SVM、DPM等）。这些方法虽然在特定场景下取得了一定的成果，但在多样化和复杂场景下表现欠佳。随着深度学习的快速发展，基于卷积神经网络（CNN）的目标检测方法逐渐成为主流。

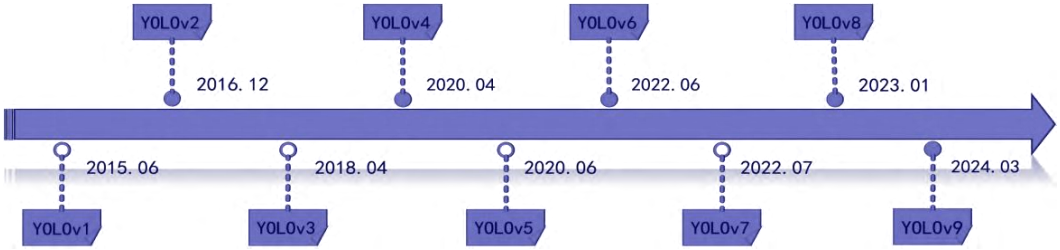


图 1-1 YOLO 算法发展线

1.3 深度学习的崛起与目标检测

深度学习的突破性进展从 AlexNet 开始，卷积神经网络被证明能够提取更深层次的特征，从而极大提升了图像分类和检测的能力。如图 1-2 所示 AlexNet 的网络结构包括八层，由五个卷积层和三个全连接层组成。在目标检测领域，两类方法逐渐形成：

（1）两阶段检测器：如R-CNN、Fast R-CNN和Faster R-CNN，它们通常分为区域生成（Region Proposal）和目标分类两步，精度较高但速度较慢。

（2）单阶段检测器：如YOLO和SSD，通过直接预测目标位置和类别实现端到端检测，具有实时性优势。[5]

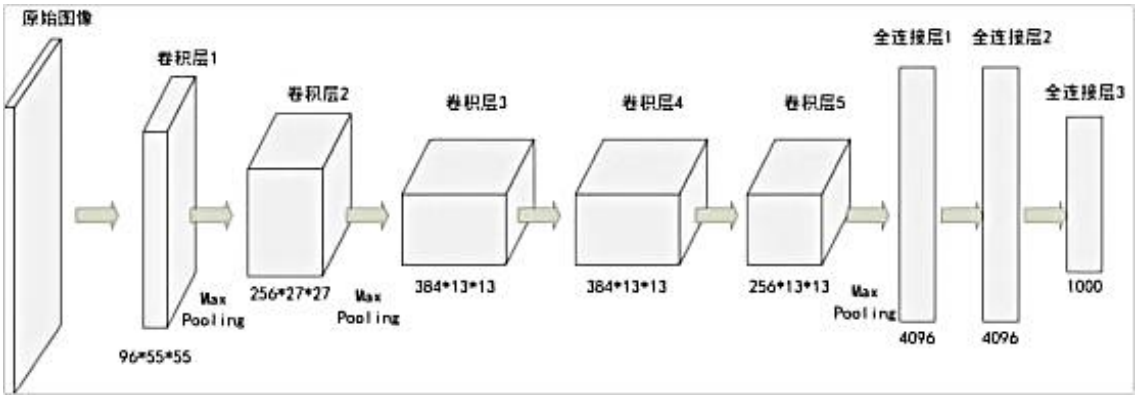


图 1-2 AlexNet 网络结构

1.4 YOLO 的意义与影响

YOLO的提出不仅提升了单阶段检测器的整体性能，还带动了相关技术（如Anchor机制、IoU Loss）的深入研究。其后续版本不断优化，不仅在检测精度和速度上超越了早期的R-CNN家族，还推动了实时目标检测在各类场景中的广泛应用。

从传统方法到深度学习，从两阶段检测器到单阶段检测器，目标检测技术的每次进步都基于特定应用场景的需求演变。YOLO通过创新的单次检测理念，成功兼顾了速度与精度的平衡，推动了目标检测技术的实际落地。与早期的网络模型如 ResNet、Faster-RCNN 等相比，YOLO 模型具有更快的检测速度和更好的检测精度。[2]

第 2 章　YOLO 系列中的目标检测技术

2.1 边界框回归与 IOU 优化

边界框回归是目标检测的核心任务之一，主要用于预测目标的空间位置和尺寸。[6]YOLO 系列通过持续优化边界框预测的损失函数和匹配机制，逐步提升了模型的检测精度和效率。

YOLO 从第一代开始即采用 IOU（Intersection Over Union） 作为评估预测框与真实框匹配程度的核心指标。IOU 的直观性和高效性使其成为目标检测中的标准工具，但其在无交集情况下梯度消失的问题也限制了优化效果。

为克服基础 IOU 的局限性，YOLO 系列逐渐引入了一系列改进：

（1）GIoU（Generalized IOU）：通过包络框惩罚项解决了无交集情况下的梯度问题。

（2）DIoU（Distance IOU）：在 GIoU 的基础上加入中心点距离约束，增强了边界框定位能力。

（3）CIoU（Complete IOU）：进一步综合长宽比一致性约束，使模型预测更加贴近真实目标。

这些改进有效提升了边界框回归的鲁棒性和模型收敛速度。

在 YOLO 系列中的演化中YOLOv1 和 YOLOv2 采用基础 IOU 进行目标匹配，表现出较高效率但对小目标较弱。YOLOv3 开始结合 Anchor 机制，提升了多尺度目标检测的适配性。YOLOv4 及之后 引入 GIoU、DIoU 和 CIoU，显著提高了边界框预测的准确性，特别是在密集目标和小目标场景中表现突出。利用YOLO算法实时处理的优势，能够优化了传统YOLO算法的锚框和网络结构。[3]

2.2 非极大值抑制（NMS）

在 YOLO 系列模型中，非极大值抑制（Non-Maximum Suppression, NMS）是目标检测后处理环节的核心技术，其作用是从众多预测框中筛选出最优框，避免重复检测。

NMS 的过程包括以下几步：

（1）置信度排序：根据模型输出的置信度对预测框降序排列。

（2）保留最高置信度框：将当前置信度最高的框作为候选框。

（3）抑制高度重叠框：对其他框与候选框计算 IOU（交并比），将 IOU 超过设定阈值的框抑制掉。

（4）重复步骤 2 和 3：直到所有框处理完毕。

这一方法通过置信度和空间重叠度的结合，有效减少了冗余框，确保每个目标只输出一个检测框。

在 YOLO 系列中的应用中 YOLOv1 和 YOLOv2 采用传统固定阈值的 NMS 方法。优点是简单高效，但在目标密集场景下，可能导致漏检。在YOLOv3 的改进中多尺度检测中分别应用 NMS，显著提升了对小目标的检测能力。[7]YOLOv4 和 DIoU-NMS 的引入里YOLOv4 引入了 DIoU（Distance IOU）作为筛选标准，在 NMS 中考虑候选框的中心点距离，优势体现为提升了密集目标场景下的检测精度。在YOLOv5 及之后版本中则采用了 Soft-NMS 方法，不直接剔除重叠框，而是根据 IOU 动态调整其置信度。优势体现为能够提高检测结果的准确性和快速、易于实现，适用于实时检测场景。局限性则表现为固定 IOU 阈值缺乏适应性，难以处理目标形状多样化的情况以及在密集目标场景中，可能出现漏检问题。

未来 NMS 的优化方向主要包括：

（1）自适应 NMS：动态调整 IOU 阈值，适应不同目标的密集程度。

（2）学习式 NMS：结合深度学习，自动学习抑制策略。

（3）端到端检测：减少候选框的生成过程，从网络中直接输出高质量框。

NMS 的持续优化，使 YOLO 系列在复杂场景下表现更加稳定，同时兼顾实时性与精度，成为其目标检测流程中的重要环节。

2.3 Focal Loss

在 YOLO 系列的目标检测中，Focal Loss 被引入以解决目标检测过程中正负样本不平衡的问题，特别是对于小目标和低置信度目标的检测优化起到了重要作用。

在目标检测任务中，类别不平衡是一个普遍问题，尤其是在正样本（目标）占比极少的情况下，背景区域（负样本）的数量占主导，传统交叉熵损失（Cross Entropy, CE）在这种情况下容易偏向负样本，导致检测性能下降。Focal Loss 是在交叉熵损失的基础上，通过增加动态调节因子重点关注难分类样本的一种优化方法。其定义为：

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

（1）p_t：预测的目标类别概率。

（2）α_t：正负样本的权重系数，用于平衡两者的贡献。

（3）γ ：聚焦因子，控制对易分类样本的损失抑制程度。

在YOLOv3 和之前版本中YOLOv3 及之前版本主要采用传统的交叉熵损失，对正负样本的不平衡问题解决有限，导致小目标和低置信度目标的检测性能欠佳。在YOLOv4 中，引入了 Focal Loss 的思想，用于处理类别预测的类别不平衡问题。在置信度预测上，Focal Loss 增强了模型对低置信度目标的关注，显著提高了小目标检测的效果。 YOLOv5 则进一步拓展了 Focal Loss 的使用范围，使其适用于置信度损失（Confidence Loss）和分类损失（Classification Loss）。 用户可通过超参数调节 α_t和γ 来适配不同的目标检测场景，如密集目标检测或复杂背景下的检测任务。

Focal Loss能够通过调整正负样本的损失权重，平衡背景与前景目标的贡献，对低置信度目标或小目标的检测精度有显著提升，提高模型在复杂场景中的适应能力。

但是局限性表现在引入动态权重后，Focal Loss 的计算量高于传统交叉熵损失。调节因子α_t 和γ 的选择对检测性能影响较大，需要针对具体任务进行优化。

通过引入 Focal Loss，YOLO 系列显著优化了类别不平衡和目标置信度的学习过程，对小目标和低置信度目标的检测性能提升尤为明显。Focal Loss 的持续优化，使其成为目标检测技术中的重要组成部分，特别是在复杂场景下表现出强大的适应能力。

2.4 多尺度检测与特征金字塔（FPN）

目标检测中的多尺度检测是提升模型性能的核心技术之一，其目标是解决目标大小差异带来的检测难度，尤其是小目标和远距离目标的检测。YOLO 系列模型通过逐步引入特征金字塔网络（Feature Pyramid Network, FPN）和多尺度预测，显著提升了对不同尺度目标的检测能力。

自然场景中的目标具有明显的尺度变化特性，小目标往往包含更少的特征信息，且易受遮挡和背景干扰；大目标尽管信息丰富，但可能因上下文信息丢失而影响检测精度。因此，模型需要在不同尺度下提取丰富的特征信息，以同时适应大目标和小目标的检测需求。

在深度学习早期阶段，经典的检测器如 R-CNN 系列仅使用单一特征层进行预测，难以兼顾多尺度目标的检测。为解决这一问题，FPN 被提出，用于从不同层次的特征图中提取多尺度信息，增强模型的检测能力。

特征金字塔网络（FPN）是一种多尺度特征融合方法，通过结合深层语义特征与浅层空间特征，构建出对多尺度目标更加敏感的特征表示。其主要思想包括：

（1）自顶向下路径： 将深层语义特征逐层传递到浅层，通过上采样操作将特征图调整至相同分辨率。

（2）横向连接： 将浅层的空间特征与深层的语义特征进行融合，从而兼顾细节信息和高级语义。

（3）多尺度预测：在每个尺度的特征图上独立进行目标检测，使模型能够同时处理大目标和小目标。如图 2-1 所示。

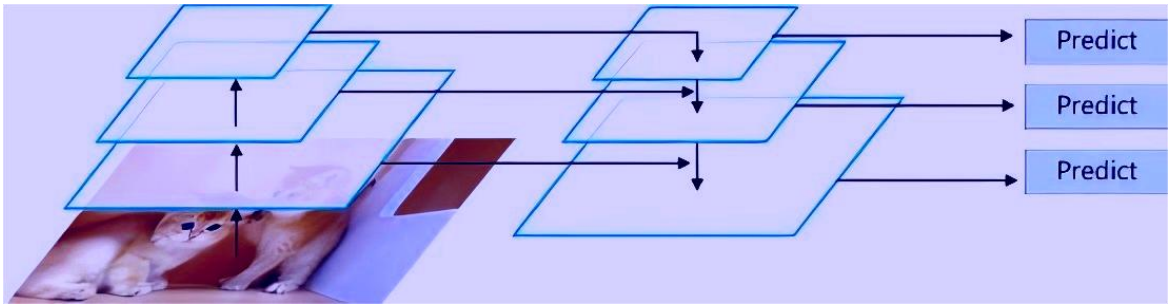


图 2-1 FPN 结构示意图

FPN 的提出极大地推动了目标检测模型在多尺度场景下的性能提升，并成为现代目标检测器的基础组件之一。

多尺度检测和特征金字塔网络（FPN）的引入，显著提升了目标检测模型对不同尺度目标的适应能力，并为小目标检测性能的提升提供了重要支持。在 YOLO 系列中，多尺度检测逐步演化，从 YOLOv3 的简单三层预测到 YOLOv4 的 PANet 改进，再到 YOLOv5 的高效特征融合机制，体现了模型对复杂检测任务的持续优化。未来，多尺度检测仍将在提高模型性能和适应性方面扮演关键角色。

第 3 章　YOLOv1——You Only Look Once

Joseph Redmon等人的YOLO发表在CVPR 2016上。它首次提出了一种实时端到端物体检测方法。YOLO代表“You Only Look Once”，指的是它能够通过一次网络完成检测任务，而之前的方法要么使用滑动窗口，然后使用需要每幅图像运行数百或数千次的分类器，要么使用更先进的方法将任务分为两步，其中第一步检测可能存在物体或区域建议的区域，第二步对建议运行分类器。此外，YOLO使用更直接的输出，仅基于回归来预测检测输出，而 Fast R-CNN则使用两个单独的输出，即概率分类和框坐标回归。

3.1 YOLOv1 工作原理

YOLOv1通过同时检测所有边界框统一了物体检测步骤。为此，YOLO将输入图像分成S×S的网格，并预测同一类的B个边界框，以及每个网格元素对C个不同类的置信度。每个边界框预测包含五个值： Pc、bx、by、bh、bw，其中Pc是框的置信度分数，反映模型对框包含物体的信心程度以及框的准确 程度。bx和by坐标是框相对于网格单元的中心，bh和bw是框相对于整幅图像的高度和宽度[8]。YOLO的输出是S×S×(B×5+C)的张量，后面可选地进行非极大值抑制(NMS)以消除重复检测。在最初的YOLO论文中，作者使用了包括20个类(C=20)的PASCAL VOC数据集 ； 7×7 的网格(S =7)，每个网格元素最多2个类 (B =2)，从而给出 7×7×30 的输出预测。图2-1显示了简化的输出向量，其中考虑了 3×3 网格、3 个类以及每个网格8个值的单个类。在这个简化的情况下，YOLO的输出为3×3×8。YOLOv1在PASCAL VOC 2007数据集上实现了63.4的平均精度（AP）。

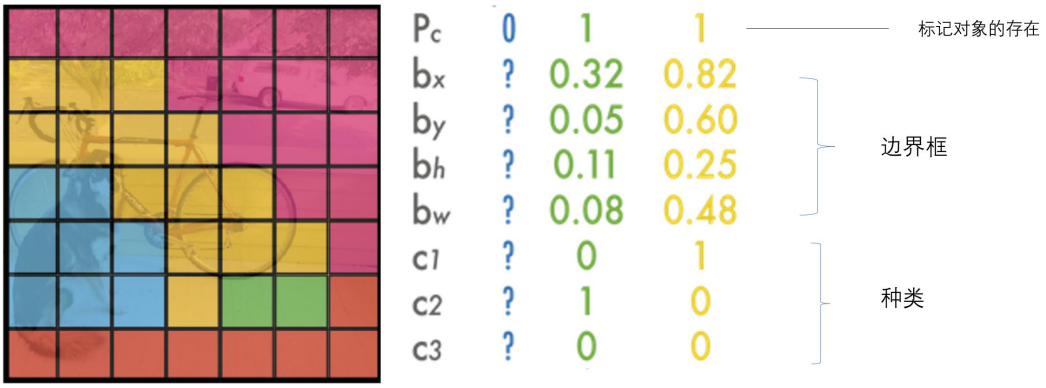


图 2-1：YOLO 输出预测。该图描绘了一个简化的 YOLO 模型，该模型具有一个三乘三网格、三个类以及每个网格元素一个类预测，以产生一个包含八个值的向量

3.2 YOLOv1 架构

YOLOv1架构包含24个卷积层，后面跟着两个全连接层，用于预测边界框坐标和概率。受到 GoogLeNet和Network in Network的启发，YOLO使用1×1卷积层，以减少特征图的数量并保持参数数量相对较低。

3.3 YOLOv1 的优势与局限性

YOLO的简单架构及其新颖的全图像一次性回归功能使其比现有的物体检测器快得多，从而实现实时性能。虽然YOLO的运行速度比任何物体检测器都要快，但与Fast R-CNN等最先进的方法相比，定位误差较大。造成这种限制的 主要原因有三个：

1.只能在网格单元中检测到最多两个同一类别的物体，这限制了预测附近物体的能力对象。

2. 很难预测训练数据中未见过的长宽比的物体。

3. 由于下采样层的局限，只能学习粗糙的物体特征。

第 4 章　YOLOv2——更好、更快、更强

YOLOv2由Joseph Redmon和Ali Farhadi在CVPR 2017上发布。能够检测 9000 个类别。改进如下：

1.所有卷积层上的批量归一化提高了收敛性，并充当了正则化器来减少 过度拟合。

2.高分辨率分类器。与YOLOv1一样，他们使用224×224的ImageNet对模型进行了预训练。但这次，他们在分辨率为448×448的ImageNet上对模型进行了10个epoch的微调，从而提高了网络在更高分辨率输入上的性能。

3.完全卷积。他们删除了密集层并使用了完全卷积架构。

4.使用锚框预测边界框。他们使用一组先验框或锚框，这些框具有预定义形状，用于匹配对象的原型形状。每个网格单元定义多个锚框，系统预测每个锚框的坐标和类别。网络输出的大小与每个网格单元的锚框数量成正比。

5.维度聚类。选择好的先验框有助于网络学习预测更准确的边界框。作者在训练边界框上运行了k-means聚类，以找到好的先验。他们选择了五个先验框，在召回率和模型复杂性之间提供了良好的权衡。

6.直接位置预测。与其他预测偏移量的方法不同，YOLOv2遵循相同的理念，预测相对于网格单元的位置坐标。网络为每个单元预测五个边界框，每个边界框有五个值tx、ty、tw、th和to，其中to相当于YOLOv1中的Pc 。

7.细粒度特征。与YOLOv1相比，YOLOv2删除了一个池化层，以便为416×416的输入图像获得13×13的输出特征图或网格。YOLOv2还使用了一个直通层，该层采用26×26×512特征图并通过将相邻特征堆叠到不同通道（而不是通过空间子采样丢失它们）对其进行重组。这会生成13×13×2048特征图，这些特征图在通道维度上与较低分辨率的13×13×1024图连接在一起，以获得13×13×3072特征图。

8.多尺度训练。由于YOLOv2不使用全连接层，因此输入可以是不同的大小。为了使YOLOv2对不同的输入大小具有鲁棒性，作者随机训练模型，每十个批次更改输入大小（从320×320到608×608）。

4.1 YOLOv2 架构

YOLOv2使用的主干架构称为Darknet-19，包含19个卷积层和5个最大池化层。与 YOLOv1的架构类似，它受到Network in Network的启发，使用1×1，3×3之间的卷积来减少参数数量。此外，如上所述，他们使用批量归一化来规范化并帮助收敛。表2显示了带有对象检测头的整个 Darknet-19主干。使用PASCAL VOC数据集时，YOLOv2预测五个边界框，每个边界框有5个值和20个类。对象分类头用一个具有1000个过滤器的单个卷积层取代最后四个卷积层，后面跟着一个全局平均池化层和一个Softmax。

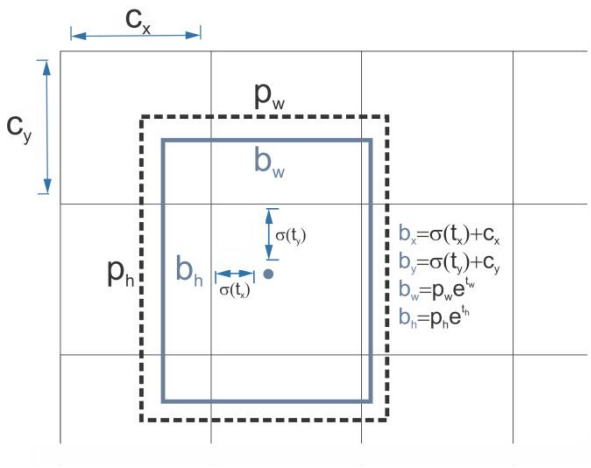


图 3-1 ：边界框预测。框的中心坐标是使用预测的 tx、ty 值通过 S 型函数并偏移网格单元 cx、cy 的位置获得的。

第 5 章　YOLOv3

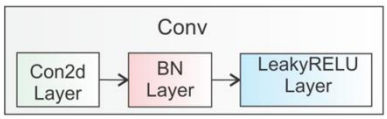
YOLOv3由Joseph Redmon和Ali Farhadi于2018年在ArXiv上发表[9]。它包含重大变化和更大的架构，以与最先进的技术保持一致，同时保持实时性能。

5.1 YOLOv3 架构

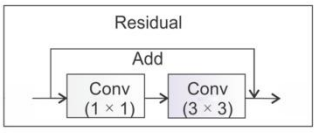
YOLOv3中提出的架构主干称为 Darknet-53。它用以下方法替换了所有最大池化层：步幅卷积并添加残差连接。总共包含53个卷积层。图4-1显示了架构细节。Darknet-53主干网络获得的Top-1和Top-5准确率与ResNet-152相当，但速度几乎快2倍。

Layer	Filters	size	Repeat		Output size
Image					416 × 416
Conv	32	3 × 3/1	1		416 × 416
Conv	64	3 × 3/2	1		208 × 208
Conv	32	1 × 1/1	Conv		208 × 208
Conv	64	3 × 3/1	Conv	×1	208 × 208
Residual			Residual		208 × 208
Conv	128	3 × 3/2	1		104 × 104
Conv	64	1 × 1/1	Conv		104 × 104
Conv	128	3 × 3/1	Conv	×2	104 × 104
Residual			Residual		104 × 104
Conv	256	3 × 3/2	1		52 × 52
Conv	128	1 × 1/1	Conv		52 × 52
Conv	256	3 × 3/1	Conv	×8	52 × 52
Residual			Residual		52 × 52
Conv	512	3 × 3/2	1		26 × 26
Conv	256	1 × 1/1	Conv		26 × 26
Conv	512	3 × 3/1	Conv	×8	26 × 26
Residual			Residual		26 × 26
Conv	1024	3 × 3/2	1		13 × 13
Conv	512	1 × 1/1	Conv		13 × 13
Conv	1024	3 × 3/1	Conv	×4	13 × 13
Residual			Residual		13 × 13

(a)卷积层参数



(b)步幅卷积



(c)残差链接

图 4-1：YOLOv3 Darknet- 53 主要结构。YOLOv3 的架构由 53 个卷积层组成，每个卷积层都具有批量归一化和 Leaky ReLU 激活。此外，残差连接将整个网络 中 1×1 卷积的输入与 3×3 卷积的输出连接起来。

5.2 YOLOv3 多尺度预测

除了更大的架构之外，YOLOv3的一个重要特性是多尺度预测，即在多个网格大小下进行预测。这有助于获得更精细的细节框，并显著改善小物体的预测，这是YOLO先前版本的主要弱点之一。图4-2所示的多尺度检测架构的工作原理如下：第一个输出标记为y1，相当于YOLOv2的输出，其中13×13的网格定义输出。第二个输出y2通过将 Darknet-53的(Res×4)之后的输出与 (Res×8)之后的输出连接起来而组成。特征图具有不同的大小，即13×13和26×26，因此在连接之前有1个上采样操作。最后，使用上采样操作，第三个输出y3将26×26特征图与52×52特征图连接起来。对于具有80

个类别的COCO数据集，每个尺度提供一个形状为N×N×[3×(4+1+80)]的输出张量，其中N×N是特征图（或网格单元）的大小，3表示每个单元的框数，4+1包括四个坐标和对象分数。

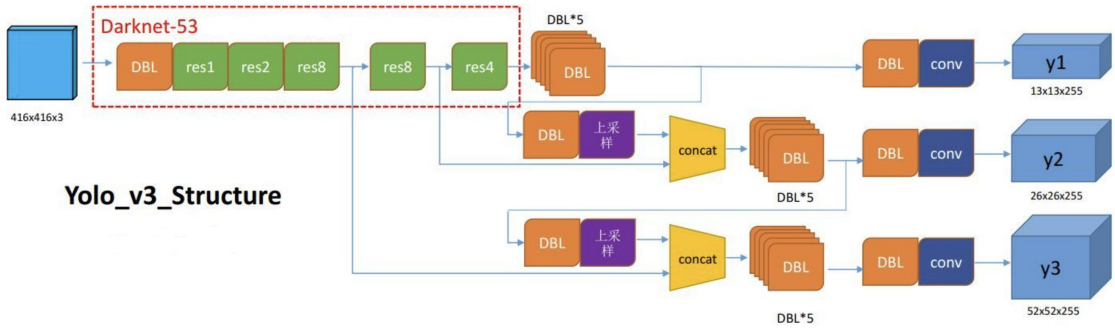


图 4-2：YOLOv3 多尺度检测架构。Darknet- 53 主干的输出分支为三个不同的输出，标记为 y1、y2 和 y3，每个输出的分辨率都有所增加。使用非极大抑制对最终预测框进行过滤。CBL（Convolution- BatchNorm- Leaky ReLU）块包括一个带有批量归一化和泄漏 ReLU 的卷积层。

第 6 章　YOLOv4

2020年4月，Alexey Bochkovskiy、Chien-Yao Wang和Hong-Yuan Mark Liao在ArXiv上发布了YOLOv4的论文[10]。YOLOv4保留了相同的YOLO理念层次和暗网框架，并且这次改进非常令人满意，于是社区迅速将此版本作为官方YOLOv4。

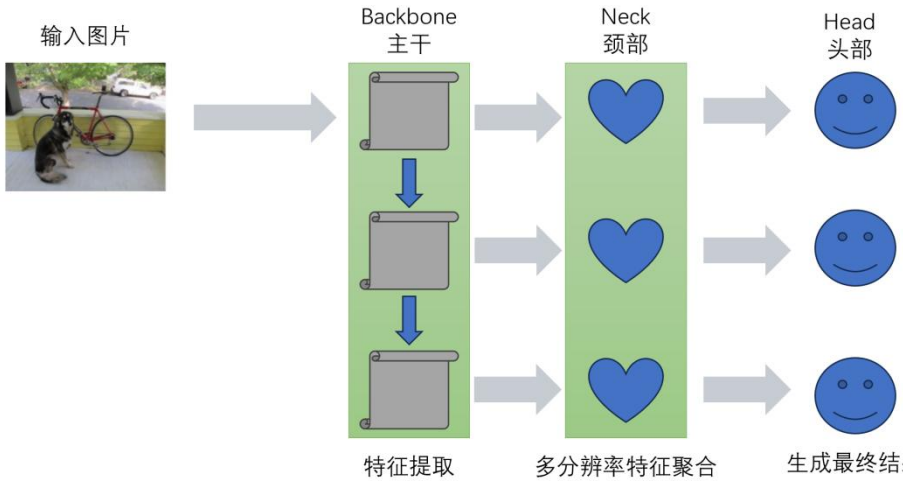


图 5-1：现代物体检测器的架构可以描述为主干、颈部和头部。主干通常是卷积神经网络(CNN)，它从不同尺度的图像中提取重要特征。颈部细化这些特征，增强空间和语义信息。最后，头部使用这些细化的特征进行物体检测预测。

YOLOv4尝试通过多种变化（分为 bag-of-freebies 和bag-of-specials）来寻找最佳平衡。bag-of-freebies是仅改变训练策略并增加训练成本但不增加推理时间的方法，最常见的是数据增强。另一方面，bag-of-specials是略微增加推理成本但显著提高准确率的方法。这些方法的例子 包括扩大感受野、组合特征和后处理等。我们总结YOLOv4的主要变化有以下几点：

具有 Bag-of-Specials (BoS) 集成的增强架构。作者尝试了多种主干架构，例如 ResNeXt50、 EfficientNet-B3和 Darknet-53。表现最佳的架构是使用跨阶段部分连接（CSPNet）和 Mish激活函数作为主干的Darknet-53的修改版（见图5-2）。对于颈部，他们使用了来自 YOLOv3-spp 的修改版空间金字塔池化（SPP）和与 YOLOv3 中的多尺度预测，但使用修改版的路径聚合网络（PANet）代替FPN，以及修改后的空间注意模块（SAM）。最后，对于检测头，他们使用与 YOLOv3相同的锚点。因此，该模型被称为 CSPDarknet53-PANet-SPP。添加到Darknet-53的跨阶段部分连接（CSP）有助于在保 持相同精度的同时减少模型的计算量。与

YOLOv3-spp中的 SPP 块一样，它可以增加感受野而不会影响推理速度。PANet 的修改版 本将特征连接起来，而不是将它们作为在原始 PANet论文中[11]。

集成 bag-of-freebies (BoF) 以实现高级训练方法。除了随机亮度、对比度、缩放、裁剪、翻转和旋转等常规增强之外，作者还实现了马赛 克增强，将四幅图像组合成一幅图像，从而可以检测其通常上下文之外的物体，同时还可以减少批量标准化对大批量大小的需求。对于正则化，他们使用了 DropBlock，它可以替代Dropout，但适用于卷积神经网络以及类标签平滑。对于检测器，他们 添加了CIoU损失和 Cross mini-bath normalization (CmBN)，用于从整个批次收集统计数据，而不是像常规批量标准化那样 从单个批量中收集统计数据。

自对抗训练 (SAT)。为了使模型对干扰具有更强的鲁棒性，对输入图像进行对抗攻击，以造成欺骗，使真实对象不在图像中，但保留原始 标签以检测正确的对象。

使用遗传算法进行超参数优化。为了找到用于训练的最佳超参数，他们在前 10% 的周期中使用遗传算法，并使用余弦退火调度程序来改变训练期间的学习率。它开始缓慢降低学习率，然后在训练过程的中途快速降低，最后略微降低。

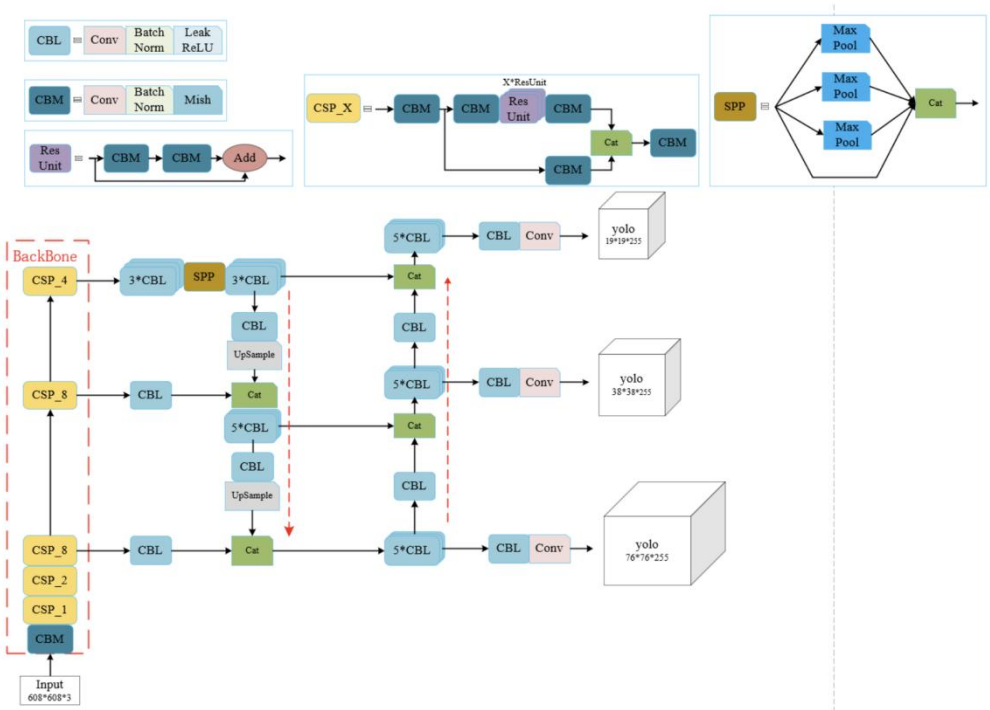


图5-2：用于对象检测的YOLOv4架构。图中的模块为CMB：卷积+批量归一化+Mish激活，CBL：卷积+批量归一化+Leaky ReLU，UP：上采样，SPP：空间金字塔池化，PANet：路径聚合网络。

第 7 章　YOLOv5

YOLOv5于2020年发布，发布的时间比YOLOv4晚几个月。YOLOv5采用了YOLOv4部分中描述的许多改进，但使用了Pytorch而不是Darknet。YOLOv5引入了一种名为自动锚框的优化算法。这个预训练工具检查并调整锚框，以确保它们适合数据集和训练设置，例如图像大小。它首先对数据集标签应用k-means算法，以生成遗传进化（GE）算法的初始条件。然后，GE算法默认经过1000代进化这些锚框，使用CIoU损失和最佳可能召回作为其适应度函数[12]。YOLOv5自发布以来推出了多种网络模型，包括：YOLOv5n、YOLOv5s、YOLOv5m、YOLON5l和YOLOv5x，其中卷积模块的宽度和深度各不相同，以适应特定的应用和硬件要求。

以YOLOv5s的模型结构为例，其网络结构如图7-1所示。

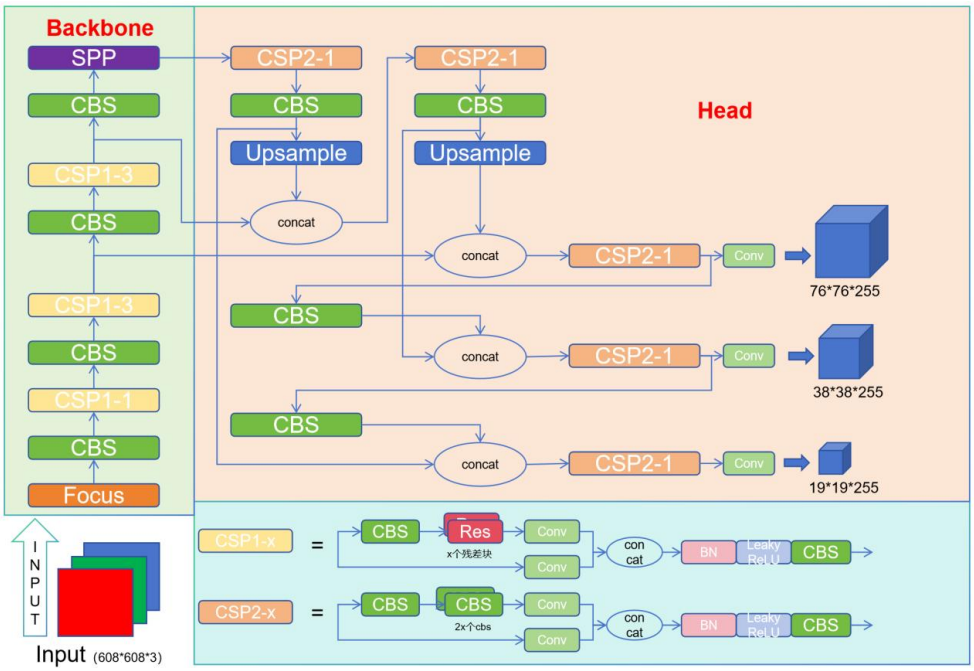


图 7-1 YOLOv5s 的网络结构

YOLOv5易于使用、训练和部署。Ultralytics提供了适用于iOS和Android的移动版本，并为标签、训练和部署提供了多种集成方案。在MS COCO2017数据集上测试时，YOLOv5x在输入图像大小为640像素，批量大小为32时的平均精度达到了50.7%。它在NVIDIAV100上能够达到200FPS的速度。此外，在使用1536像素的较大输入大小并增加测试时间时，YOLOv5的平均精度达到了55.8%。

第 8 章　YOLOv6

YOLOv6于2022年9月由美团视觉AI部门发布。该网络设计包括一个高效的主干网络，使用RepVGG或CSPStackRep块，以及一个具有 PAN 拓扑结构的颈部和高效的解耦头部，采用混合通道策略。此外，论文还引入了增强的量化技术，通过后训练量化和通道级蒸馏，提升了检测器的速度和准确性[13]。总体而言YOLOv6在准确性和速度指标上超越了之前的最先进模型，如 YOLOv5、YOLOX 和 PP-YOLOE。

YOLOv6引入了一种新的主干网络，称为EfficientRep，该网络基于RepVGG[14]，具有比之前的YOLO主干更高的并行性。在颈部部分，YOLOv6使用了增强的PAN（路径聚合网络），结合了RepBlocks或CSPStackRep块，以适应更大的模型。此外，继承了YOLOX的设计理念，YOLOv6还开发了一个高效的解耦头部。这些改进使得YOLOv6 在性能上得以显著提升。

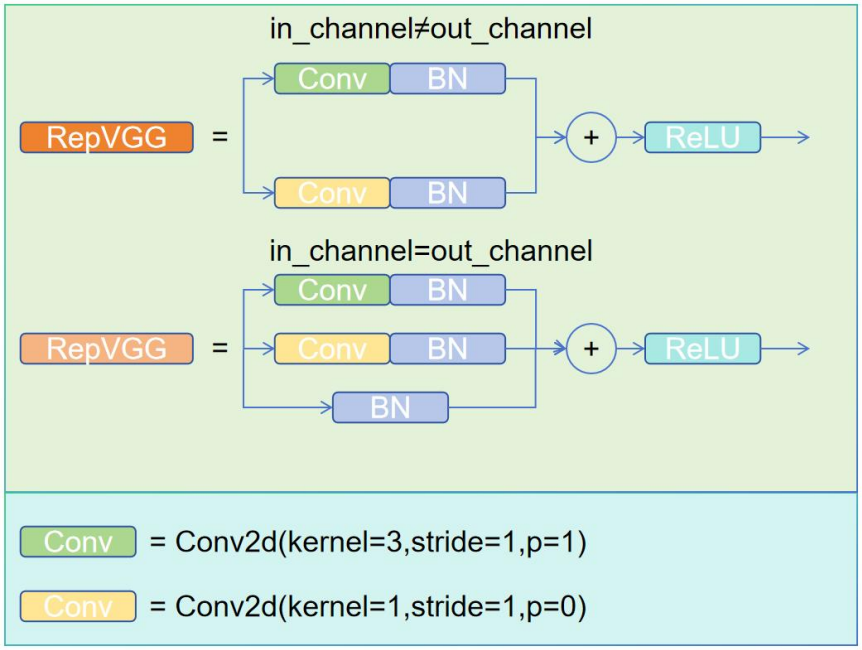


图 8-1 YOLOv6 中使用的 RepVGG 模块

YOLOv6的作者提供了八个不同规模的模型，从YOLOv6-N到YOLOv6-L6。在MS COCO2017数据集测试中，最大的模型在NVIDIA Tesla T4上达到了约29 FPS的速度，并实现了57.2%的平均精度。这些不同规模的模型使得YOLOv6能够在多种应用场景中灵活使用。

第 9 章　YOLOv7

9.1 YOLOv7 的网络结构

与YOLOv4相似，YOLOv7仅使用MS COCO2017数据集进行训练，没有使用预训练的主干网络。YOLOv7 提出了若干架构改进和一系列“免费策略”，这些改进提高了准确性而不影响推理速度，仅增加了训练时间。在具体的网络模块上沿用了YOLOv5的模块实现，并提出了部分新的模块，YOLOv7的网络结构如图9-1所示。

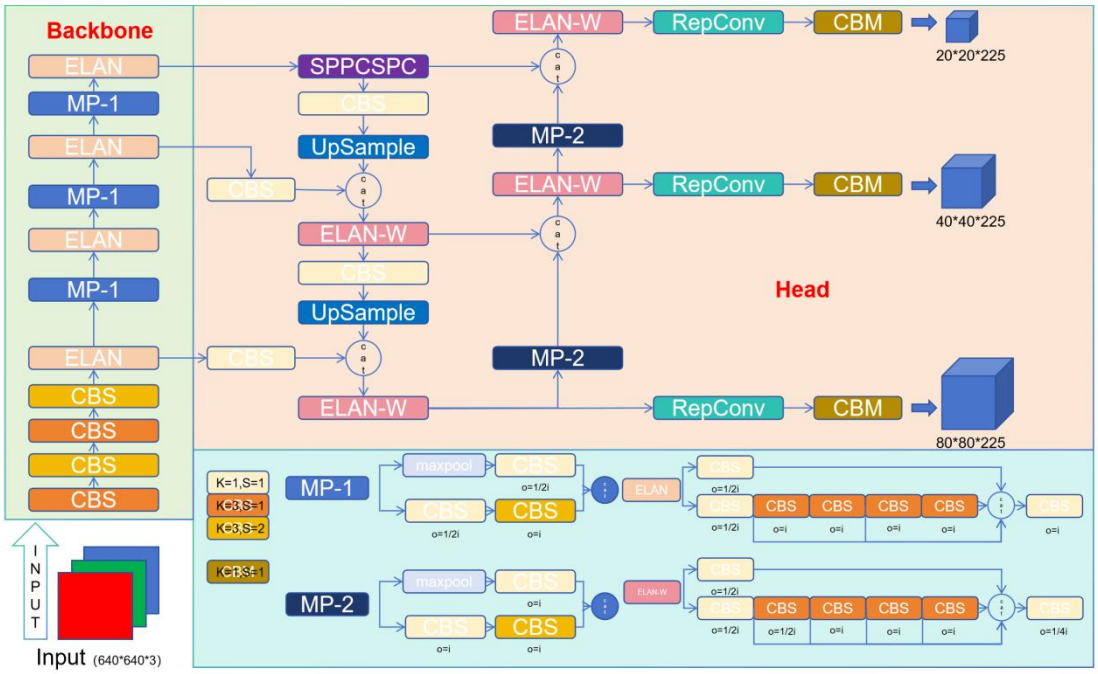


图 9-1 YOLOv7 的网络结构，其中包含了新实现的网络模块 MP-2,ELAN-W。

9.2 扩展的高效聚合网络

YOLOv7中提出了基于ELAN的扩展版本——E-ELAN，如图9-2所示。其主要架构展示了在大规模ELAN中，无论梯度路径长度和计算模块堆叠数量如何，系统都能达到稳定状态。无限堆叠更多计算模块可能会破坏这一稳定状态并降低参数利用率。E-ELAN通过扩展、混洗和合并来增强网络的学习能力，而不破坏原有的梯度路径。

在架构设计上，E-ELAN仅对计算模块的架构进行修改，而过渡层的架构保持不变。该策略采用组卷积来扩展计算模块的通道和基数，并在每个计算层的所有计算模块中应用相同的组参数和通道乘数。每个计算模块计算出的特征图会根据设置的组参数g被混洗成g组，然后进行拼接，确保每组特征图的通道数与原始架构相同。

最后，通过合并基数的方法将g组特征图相加。E-ELAN不仅维护了原有ELAN的设计架构，还能引导不同组的计算模块学习到更丰富的特征。

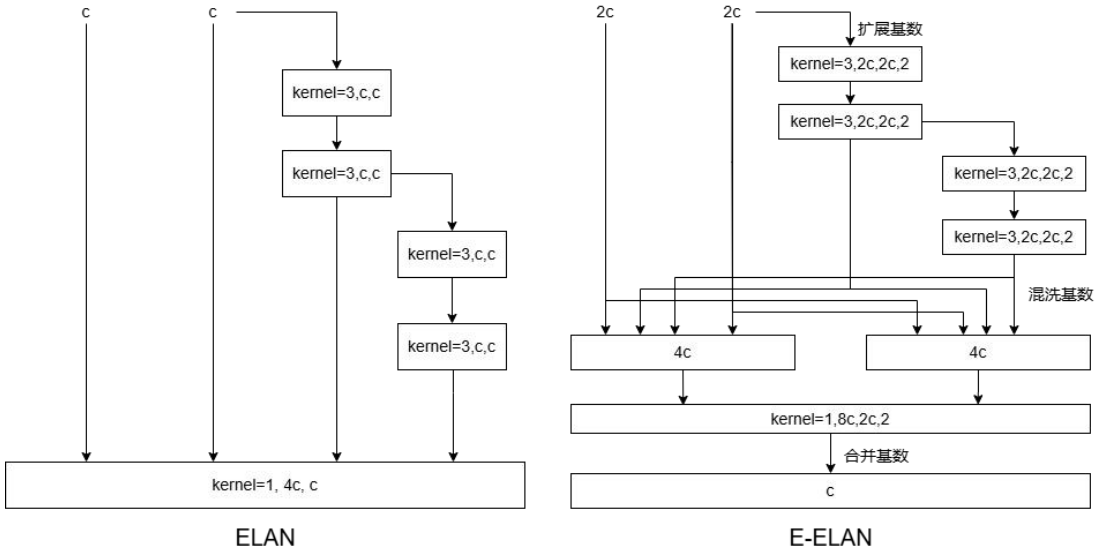


图 9-2 ELAN 网络结构和 E-ELAN 网络结构

9.3 可训练的 bag-of-freebies

除了网络结构上的改进以外，YOLOv7中还提出了多种可以提高目标检测精度的优化方法，作者将它们称为bag-of-freebies，这部分会介绍主要的优化模块和优化方法[15]。

9.3.1 卷积重新参数化

卷积重新参数化最初由RepVGG提出，沿用ResNet的残差边，在每一层都使用了残差连接，3x3卷积接上一个恒等连接和1x1的卷积。RepVGG是一种在训练时具有多分支拓扑结构，但在推理时可以等效融合为单个3x3卷积的可重参数化结构。这种设计使得模型能够在训练时通过多分支学习不同的特征表示，从而提高模型的性能。**错误！未定义书签。**YOLOv6中使用了RepConv模块提升模型性能。RepConv在VGG结构中应用时能够显著提升模型表现，但当将其应用于ResNet和DenseNet时，RepConv中的恒等连接可能会干扰残差学习和特征拼接，从而导致模型准确性显著下降。在YOLOv7中，引入了梯度流传播路径的分析方法，以探讨如何将重新参数化的卷积有效地结合到不同网络架构中，并设计了一种新的卷积重参数化方法——RepConvN，即不使用恒等连接的RepConv。RepConv和RepConvN的结构化表示如图9-3所示。

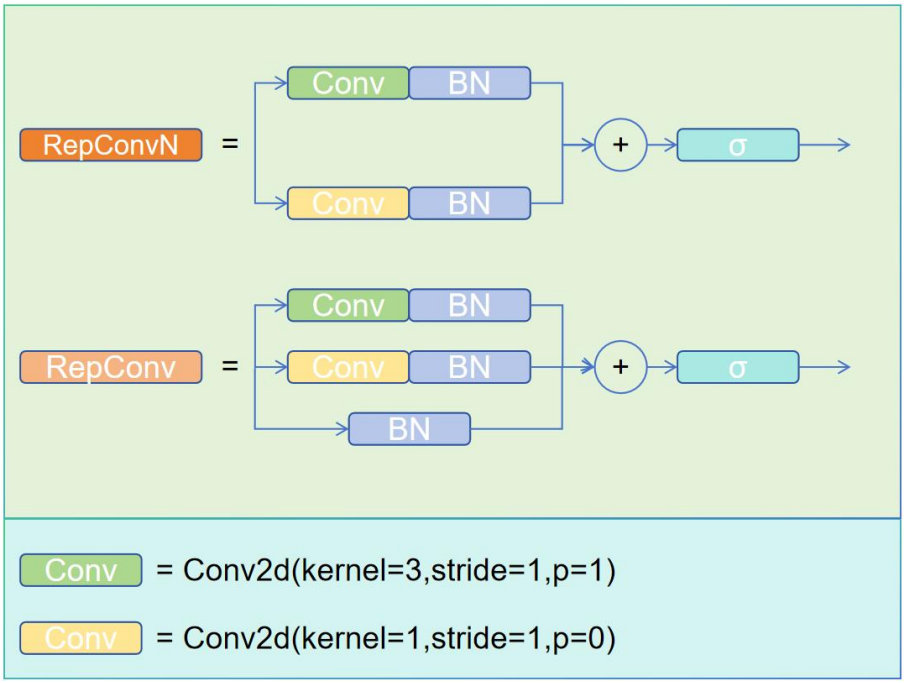


图 9-3 使用恒等连接的 RepConv 和不使用恒等连接的 RepConvN

在YOLOv7中，作者通过不同的网络连接结构测试，表明RepConvN相比于RepConv，可以更好地与ResNet等网络思想结合，如图9-4所示。

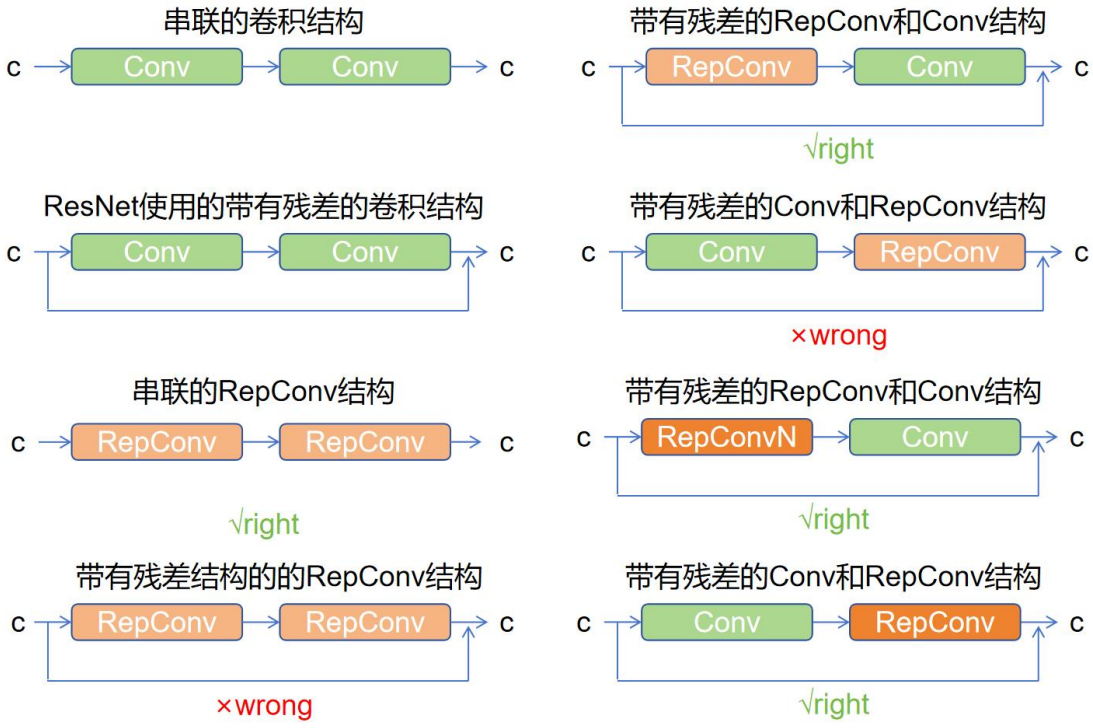


图 9-4 RepConv 和 RepConvN 在不同结构中的表现，RepConvN 可以比 RepConv 更好地应用在带有残差的网络结构中。

在RepConvN中，当一个具有残差连接或特征串联的卷积层被重新参数化的卷积替代时，必须避免使用恒等连接。这一设计考虑旨在确保网络的学习能力和特征表达不被削弱，从而在保留网络原有优势的基础上，进一步优化性能。

9.3.2 一种粗到细的标签分配策略

YOLOv7中还提出了一种优化的标签分配方法，借助深度监督的方法，添加一个辅助头辅助训练，由指导头进行最终的输出。当前的主流方法采用独立的辅助头和指导头，利用它们各自的预测结果和真实情况来执行标签分配。YOLOv7提出了一种新方法——指导头导向标签分配器，主要基于引导头的预测结果和真实值来计算，并通过优化过程生成软标签。这组软标签将用作辅助头和引导头的训练。这样做的原因是因为指导头具有相对较强的学习能力，因此由其生成的软标签应更能代表源数据和目标之间的分布和相关性。此外，可以将这种学习视为一种广义残差学习。通过让较浅的辅助头直接学习指导头已经学习的信息，指导头将更能够专注于学习尚未学习的残差信息。在此基础上，YOLOv7又提出了从粗到细的标签分配策略，细标签与引导头在标签分配器上生成的软标签相同，粗标签是通过放宽认定正类目标的条件生成的，即允许更多的网格作为正类目标。这两种标签分配策略如图9-5所示。

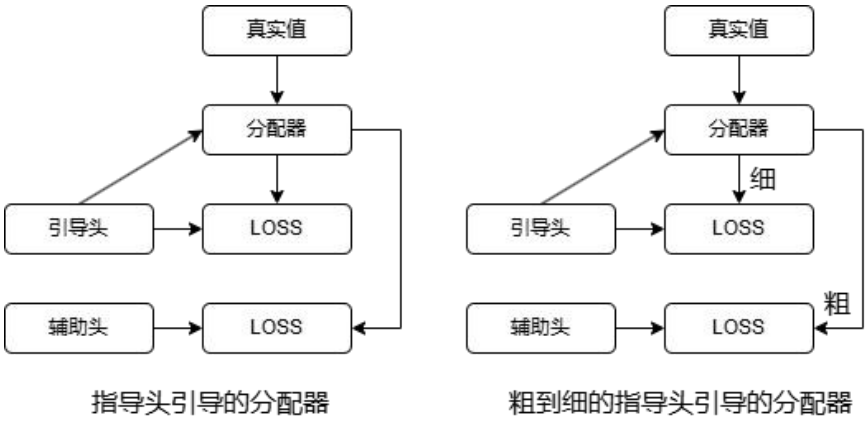


图 9-5 指导头导向的标签分配器和从粗到细的标签分配

第 10 章　YOLOv8

YOLOv8[16]于2023年1月由开发YOLOv5的Ultralytics公司发布。YOLOv8供五个比例版本：YOLOv8n（微型），YOLOv8s（小型），YOLOv8m（中型），YOLOv8l（大型）和YOLOv8x（特大型）。YOLOv8支持多种视觉任务，如目标检测、分割、姿态估计、跟踪和分类。

10.1 YOLOv8 架构

图1-1显示了YOLOv8的详细架构。YOLOv8使用了与YOLOv5类似的主干，只是在CSPLayer上做了一些更改，现在称为C2f模块。C2f模块将高层特征与上下文信息相结合，以提高检测精度。

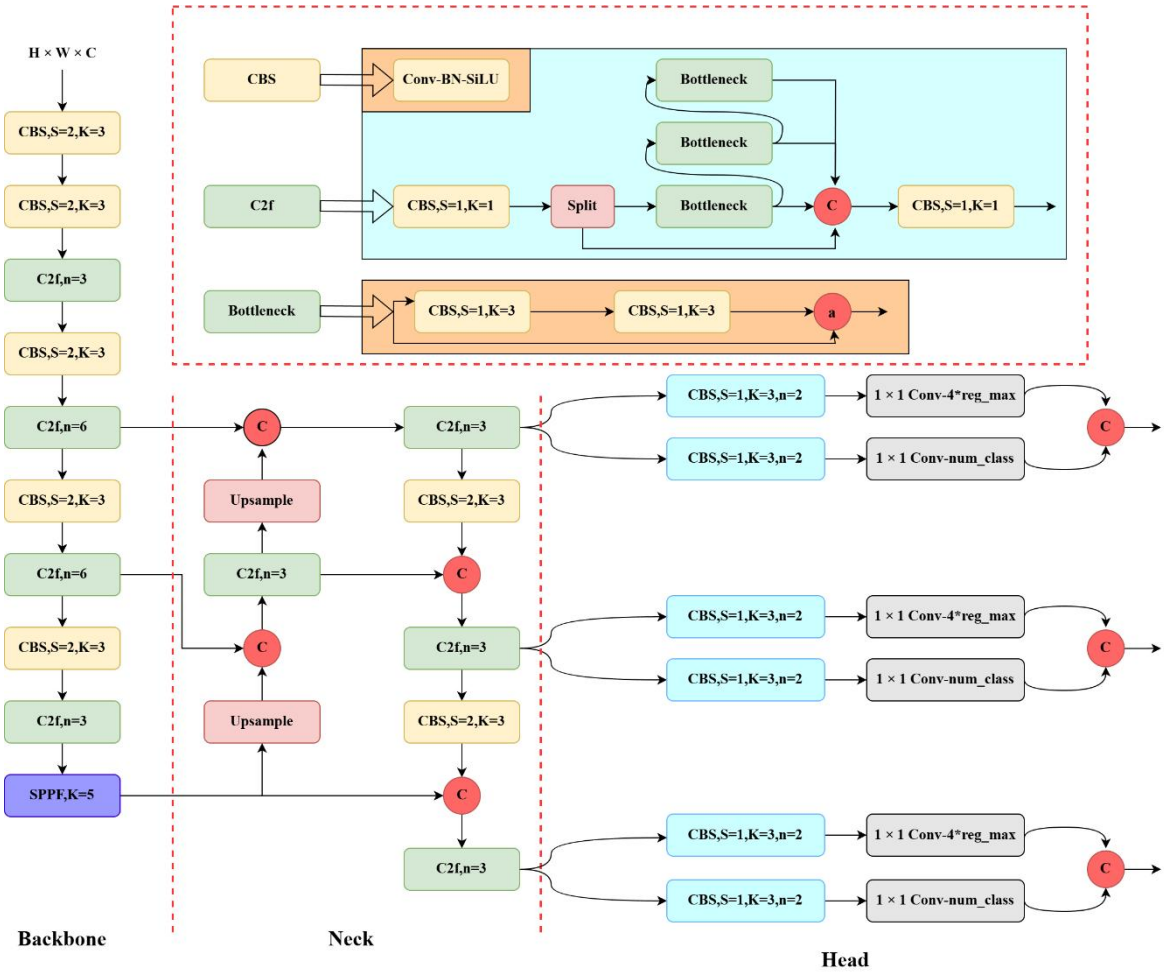


图 1-1 YOLOv8 架构

YOLOv8使用一个头部解耦的无锚点模型来独立处理似物性、分类和回归任务。这种设计允许每个分支专注于自己的任务，并提高模型的整体准确性。在YOLOv8的输出层中，他们使用sigmoid函数作为似物性分数的激活函数，表示边界框包含对象的概率。它使用softmax函数来计算类别概率，表示对象属于每个可能类别的概率。

YOLOv8使用CIoU[17]和DFL[18]损失函数来计算边界框损失，使用二进制交叉熵来计算分类损失。这些损失提高了目标检测性能，特别是在处理较小的目标时。

YOLOv8还提供了一个语义分割模型，称为YOLOv8- seg模型。其主干是CSPDarknet53特征提取器，其次是C2f模块，而不是传统的YOLO颈部架构。C2f模块后面是两个分割头，它们学习预测输入图像的语义分割掩码。该模型具有与YOLOv8相似的检测头，由五个检测模块和一个预测层组成。YOLOv8-Seg模型在各种目标检测和语义分割基准上取得了最先进的结果，同时保持了高速和效率。

YOLOv8可以从命令行界面（CLI）运行，也可以作为PIP包安装。此外，它还具有用于标记、训练和部署的多个集成。在MS COCO dataset test-dev 2017上进行评估，YOLOv8x在640像素大小的图像上取得了53.9%的AP，在NVIDIA A100和TensorRT上的速度为280 FPS。

10.2 YOLOv8 效果

以MS COCO dataset test-dev 2017为基准，YOLOv8x取得了53.9%的AP值，图像大小为640像素，超过了YOLOv5（与YOLOv5在相同输入大小下的50.7%相比）。YOLOv8在使用TensorRT的NVIDIA A100上实现了280 FPS的显著速度，它在实时应用中的效率也非常高。

第 11 章　YOLOv9

YOLOv9[19]由中国台湾 Academia Sinica、台北科技大学等机构联合开发。YOLOv9 建立在YOLOv7版本之上，融合了深度学习技术和架构设计的进步，以在对象检测任务中实现卓越的性能。YOLOv9将可编程梯度信息（PGI）概念与通用 ELAN （GELAN）架构相结合而开发，代表了准确性、速度和效率方面的重大飞跃。

11.1 YOLOv9 架构

YOLOv9提出了一种新的辅助监督框架：可编程梯度信息PGI（Programmable Gradient Information）。 PGI 主要包括三个部分，即主分支、辅助可逆分支和多级辅助信息。

（1）PGI 的推理过程仅使用了主分支，因此不需要额外的推理成本；

（2）辅助可逆分支是为了处理神经网络加深带来的问题， 网络加深会造成信息瓶颈，导致损失函数无法生成可靠的梯度；

（3）多级辅助信息旨在处理深度监督带来的误差累积问题，特别是多个预测分支的架构和轻量级模型。

YOLOv9提出了一个新的网络架构 GELAN，如图2-1所示，具体而言，研究者把 CSPNet[20]、 ELAN[21]这两种神经网络架构结合起来，从而设计出兼顾轻量级、推理速度和准确性的通用高效层聚合网络 GELAN （ Generalized Efficient Layer Aggregation Network）。研究者将最初仅使用卷积层堆叠的 ELAN 的功能泛化到可以使用任何计算块的新架构。

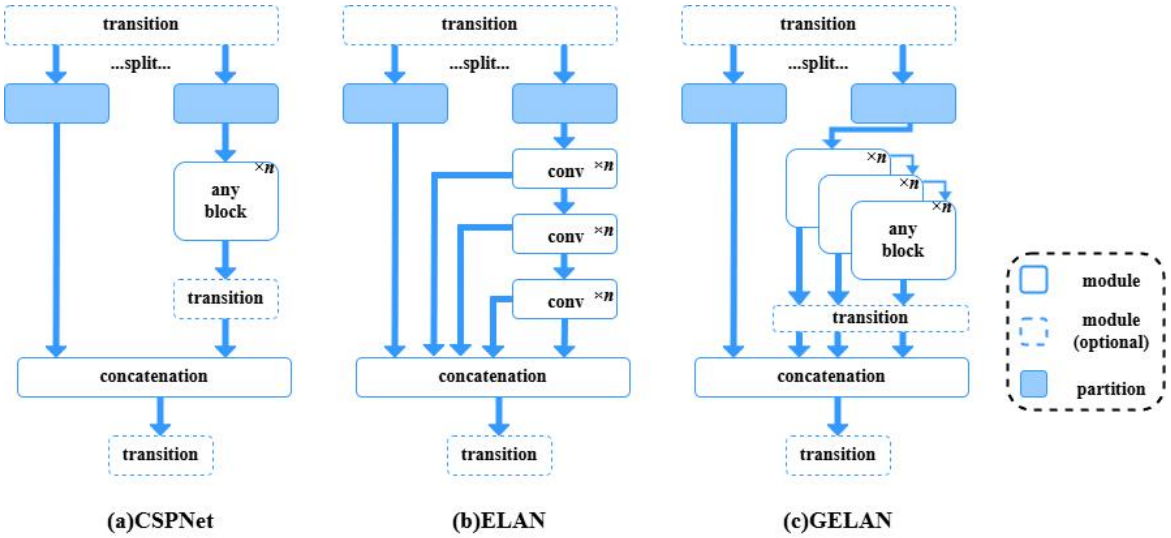


图 2-1 GELAN 架构

11.2 YOLOv9 效果

实验结果表明，YOLOv9 在 MS COCO 等基准数据集上的目标检测任务中实现了最佳性能。它在准确性、速度和整体性能方面超越了现有的实时物体检测器，使其成为需要物体检测功能的各种应用的最先进的解决方案。

另外还将 ImageNet [22]预训练模型纳入比较中，值得注意的是，使用传统卷积的 YOLOv9 在参数利用率上甚至比使用深度卷积的 YOLO MS 还要好。

11.3 YOLOv9 变体

YOLOv9t: YOLOv9变体中最小的模型，YOLOv9t专为高度受限的环境，如物联网设备和边缘计算应用程序。只有200万个参数和7.7 MB的模型大小，在速度和精度之间取得了最佳平衡，使其适用于计算资源有限的实时推理。

YOLOv9s：这个轻量级模型具有720万个参数，性能有所提高。它适用于对计算效率和精度要求不高的应用。YOLOv9s提供46.8%的mAP，同时保持26.7 MB的小模型大小，使其成为移动和嵌入式系统的可行选择。

YOLOv9m：中间层模型YOLOv9m具有2010万个参数，非常适合要求更高精度而又不牺牲实时性能的任务。它的mAP为51.4%，在一系列需要平衡精度和资源效率的目标检测场景中表现出色。

YOLOv9c：具有2550万个参数和增强的架构，YOLOv9c进行了优化，以提高精度，同时减少计算需求。与YOLOv7 AF相比，它在达到相当精度的同时，所需参数减少42%，计算能力减少21%。

YOLOv9e: YOLOv9系列中最大、最强大的模型，YOLOv9e包含5810万个参数。它提供了最高的准确性，mAP为55.6%，适用于精度至关重要的任务，如监视或医疗成像。尽管精度很高，但它仍然具有计算效率，与YOLOv8-X相比，参数减少了15%，计算量减少了27%。

第 12 章　YOLOv10

YOLOv10[23]于2024年5月由清华大学发布。YOLOv10的目标是从后处理和模型架构两个方面进一步推进YOLO的性能效率边界,这带来了竞争力的性能和低推理延迟。此外，从效率和准确性两个角度全面优化了YOLO的各个组件，这大大降低了计算开销，提高了性能，用于实时端到端对象检测。大量的实验表明，YOLOv10在各种模型尺度上都达到了最先进的性能和效率。

12.1 YOLOv10 架构

当前YOLOv1~v9仍存在一些不足，一方面，后处理对非极大值抑制（NMS）的依赖阻碍了 YOLO 的端到端部署，并对推理延迟产生了不利影响。另一方面，YOLO 中各个组件的设计缺乏全面和彻底的检查，导致明显的计算冗余，限制了模型的能力。因此YOLOv10从后处理和模型架构方面进一步提升了 YOLO 的性能-效率边界。

首先提出了 YOLO 无 NMS 训练的一致双重分配，在训练过程中将一对多和一对一策略结合起来，以确保丰富的监督和高效的端到端部署。一致匹配度量使两种策略之间的监督保持一致，从而提高了推理过程中的预测质量。

此外，YOLOv10提出了整体效率-精度驱动的模型设计策略，从效率和准确率两个角度全面优化 YOLO 的各个组件：

（1）提高效率

- 轻量级分类头：通过使用深度可分离卷积，减少分类头的计算开销。
- 空间信道解耦向下采样：将空间缩减与信道调制解耦，最大限度地减少信息损失和计算成本。
- 梯级引导程序块设计：根据固有阶段冗余调整模块设计，确保参数的最佳利用。

（2）精度提升

- 大核卷积扩大感受野，增强特征提取能力。
- 部分自我关注（PSA）：纳入自我关注模块，以最小的开销改进全局表征学习。

12.2 YOLOv10 效果

YOLOv10 提供了多个模型规模（N、S、M、B、L、X），允许用户根据性能和资源约束选择最适合的模型，具体如表3-1所示。这种可扩展性确保了 YOLOv10 能够有效应用于各种实时检测任务，从移动设备上的轻量级应用到需要高精度的复杂任务。

表 3-1 YOLOv10 型号

模型	输入尺寸	参数	AP^{val}	FLOPs	延迟
YOLOv10-N	640	2.3M	38.5%	6.7G	1.84ms
YOLOv10-S	640	7.2M	46.3%	21.6G	2.49mx
YOLOv10-M	640	15.4M	51.1%	59.1G	4.74ms
YOLOv10-B	640	19.1M	52.5%	92.0G	5.74ms
YOLOv10-L	640	24.4M	53.2%	120.3G	7.28ms
YOLOv10-X	640	29.5M	54.4%	160.4G	10.70ms

在准确性和效率方面，YOLOv10 优于YOLO 以前的版本和其他最先进的模型。例如，在 COCO 数据集上，YOLOv10-S 的速度是RT-DETR-R18 的 1.8 倍，而 YOLOv10-B 与 YOLOv9-C 相比，在性能相同的情况下，延迟减少了 46%，参数减少了 25%。

第 13 章　应用与总结

13.1 YOLO 的应用

尽管 YOLO 系列在目标检测领域的速度与准确性平衡方面处于领先地位，但其主要工作仍集中在计算机端。在进一步研究中，如何使 YOLO 更轻、更快值得深入思考，尤其是针对 Nvidia Jetson Nano 和 Raspberry Pi 等嵌入式设备。此外，将深度神经网络（DNN）模型部署在 FPGA 上以提高模型运行效率将成为一种趋势。而且，结合人工智能和物联网技术也将为人类生活带来更多便利。最后，自 YOLOv4 发布以来，集成各种先进算法已成为开发 YOLO 算法的重要途径。随着 YOLO 框架的发展，YOLO 变得更加多功能和强大，并将获得更广泛的应用。

YOLO 模型已在工业、人工智能物联网（AIoT）、医疗保健和文化遗产保护等领域广泛应用。在工业领域，许多 YOLO 应用集中在机器人、交通和个人防护设备方面。针对空盘回收机器人，Yue 等人[24]提出了一种基于 YOLOX 的轻量化盘子检测模型，解决了传统物体检测模型需存储参数的问题。Ge 等人[25]通过轻量化的 YOLO - GG 模型实现了高效的空盘回收。在交通方面，He 等人[26]设计了一种灵活高效的一阶段物体检测网络 FE - YOLO，专门用于轨道交通场景。Li 等人[27]对 YOLOv5 模型进行优化，提出了一种安全帽检测系统，保障工人安全。

在 AIoT 领域也备受关注，AIoT 是人工智能（AI）技术与物联网（IoT）融合的应用。Morioka 等人提出了一种基于 YOLO 模型的安卓系统用于古籍文本识别，该系统通过与服务器通信实现 AI 模型的识别。基于人工智能和大数据构建智能城市交通管理系统已成为一种趋势，Liu 等人利用 YOLOv3 进行车辆检测并通过图像处理估算车辆长度。

YOLO 在医疗保健领域也有广泛应用。有研究提出了一种基于 YOLOv4 - tiny 的系统，该系统在 COVID - 19 初期于 Jetson Nano 上部署，能够识别口罩佩戴状态并测量社交距离，有效保护人们的健康。Zhuang 等人将 YOLO 与改进的二维连续方程相结合，进行心脏向量流动映射（VFM）分析与评估。

此外，古籍承载着大量信息，解码这些经典书籍对于研究历史、政治和文化意义重大。Liu 和 Fujikawa [28]利用 YOLO 检测和识别甲骨文。

总之，YOLO 模型在各个领域发挥了重要作用，优化 YOLO 模型以适应不同应用已成为一种趋势。

13.2 总结

YOLO 系列目标检测模型始终致力于平衡速度与准确性，旨在确保实时性能的同时不降低检测结果的质量。在 YOLO 框架的不同版本演变过程中，这一平衡问题反复出现，每个版本都尝试以各异的方式对这两个相互竞争的目标进行优化。

在原始 YOLO 模型中，高速目标检测是其主要关注点。该模型运用单一卷积神经网络（CNN）直接从输入图像预测目标的位置和类别，以此实现实时处理。然而，这种对速度的侧重在处理小物体或重叠边界框时，导致了准确性有所下降。

后续的 YOLO 版本针对这些局限引入了改进和增强措施，同时维持了框架的实时处理能力。例如，YOLOv2 引入锚框和透传层，改进了目标定位，从而提高了准确性。此外，YOLOv3 采用多尺度特征提取架构增强了模型性能，使其能在不同尺度下更好地进行目标检测。

随着 YOLO 框架的不断发展，速度与准确性之间的权衡变得更为精细。像 YOLOv4 和 YOLOv5 等模型引入了诸如新的网络骨干、改进的数据增强技术和优化的训练策略等创新点。这些发展在不显著影响模型实时性能的情况下，大幅提高了准确性。

从 YOLOv5 开始，所有官方 YOLO 模型都对速度与准确性的权衡进行了微调，提供了不同的模型规模以满足特定应用和硬件需求。例如，这些版本通常会提供针对边缘设备优化的轻量级模型，牺牲一定的准确性以降低计算复杂度并加快处理速度。

随着 YOLO 框架的持续演进，我们预计以下趋势和可能性将对其未来发展产生影响：

一是最新技术的融合。研究人员和开发者将持续借助深度学习、数据增强和训练技术中的前沿方法改进 YOLO 架构。这种持续创新有望提升模型的性能、鲁棒性和效率。

二是基准评估的演变。当前用于评估目标检测模型的 COCO 2017 基准，可能最终会被更先进且更具挑战性的基准取代。这体现了从 YOLO 前两个版本使用的

VOC 2007 基准的转变，表明随着模型愈发复杂和精确，对更严格基准的需求也在增加。

三是 YOLO 模型和应用数量的增多。随着 YOLO 框架的发展，预计每年发布的 YOLO 模型数量会增加，其应用领域也将相应拓展。随着框架功能日益强大，它将在更广泛的领域得到应用，从家电设备到自动驾驶汽车。

四是向新领域的扩展。YOLO 模型有潜力突破目标检测和分割领域，探索如视频中的目标跟踪和 3D 关键点估计等领域。我们预计 YOLO 模型将向多模态框架发展，融合视觉与语言、视频和声音处理。随着这些模型的发展，它们可能成为满足更广泛计算机视觉和多媒体任务的创新解决方案的基础。

五是对多样化硬件的适应。YOLO 模型将进一步覆盖从物联网设备到高性能计算集群等硬件平台。这种适应性将使 YOLO 模型能依据应用的需求和限制在各种场景下部署。此外，通过根据不同硬件规格定制模型，YOLO 可为更多用户和行业提供可获取且有效的解决方案。

参考文献

[1]Yutong LI,Miao MA,Shichang LIU,Chao YAO,Longjiang GUO.YOLO-Drone:A Scale-Aware Detector for Drone Vision[J].Chinese Journal of Electronics,2024,33(04):1034-1045.

[2]LIU Qing,WU Ting-ting,DENG Ya-hong,LIU Zhi-heng.Intelligent identification of landslides in loess areas based on the improved YOLO algorithm: a case study of loess landslides in Baoji City[J].Journal of Mountain Science,2023,20(11):3343-3359.

[3]LUO Yasong,XU Jianghu,FENG Chengxu,ZHANG Kun.An accurate detection algorithm for time backtracked projectile-induced water columns based on the improved YOLO network[J].Journal of Systems Engineering and Electronics,2023,34(04):981-991.

[4]许月.基于多尺度特征融合的深度学习目标检测技术研究[D]. 安徽工程大学, 2023.

[5]张亚西.基于视频的目标检测算法研究[D]. 上海师范大学, 2020.

[6]陈灏然.基于卷积神经网络的小目标检测算法研究[D]. 江南大学, 2021.

[7]金亚飞.复杂场景下高效率车牌识别研究[D]. 安徽大学, 2019.

[8]K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

[9]J. Redmon and A. Farhadi, “Yolov3:An incremental improvement,” arXiv preprint arXiv:1804.02767, 2018.

[10]A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” arXiv preprint arXiv:2004.10934, 2020.

[11]K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[12]X. Zhu, S. Lyu, X. Wang and Q. Zhao, TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios,[C]2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 2021, pp. 2778-2788

[13]Chuyi Li, LuLu Li, Hongliang Jiang, YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications[J].arXiv preprint arXiv:2209.02976

[14]Ding X , Zhang X , Ma N ,et al.RepVGG: Making VGG-style ConvNets Great Again[J]. 2021.DOI:10.1109/CVPR46437.2021.01352.

[15]C. -Y. Wang, A. Bochkovskiy and H. -Y. M. Liao, YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors[C] //2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 7464-7475

[16]G. Jocher, A. Chaurasia, and J. Qiu, “YOLO by Ultralytics.” https://github.com/ultralytics/, ultralytics, 2023. Accessed: February 30, 2023.

[17]Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-iou loss: Faster and better learning for bounding box regression,” in Proceedings of the AAAI conference on artificial intelligence, vol. 34, pp. 12993–13000, 2020.

[18]X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, “Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection,” Advances in Neural Information Processing Systems, vol. 33, pp. 21002–21012, 2020.

[19]Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//European Conference on Computer Vision. Springer, Cham, 2025: 1-21.

[20]Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 390–391, 2020.

[21]Chien-Yao Wang, Hong-Yuan Mark Liao, and I-Hau Yeh. Designing network design strategies through gradient path analysis. Journal of Information Science and Engineering (JISE), 39(4):975–995, 2023.

[22]Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[23]Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. arXiv preprint arXiv:2405.14458, 2024.

[24]X. Yue, H. Li, M. Shimizu, S. Kawamura, L. Meng, Yolo-gd: A deep learning-based object detection algorithm for empty-dish recycling robots, Machines 10 (2022). doi:10.3390/machines10050294.

[25]X. Yue, H. Li, L. Meng, An ultralightweight object detection network for empty-dish recycling robots, IEEE Transactions on Instrumentation and Measurement 72 (2023). doi:10.1109/TIM.2023.3241078.

[26]Y. Ge, X. Yue, L. Meng, A high-efficiency dirty-egg detection system based on yolov4 and tensorrt, in: 2022 International Conference on Advanced Mechatronic Systems (ICAMechS), Toyama, Japan, 2022, pp. 59–63.

[27]D. He, Z. Zou, Y. Chen, B. Liu, J. Miao, Rail transit obstacle detection based on improved cnn, IEEE Transactions on Instrumentation and Measurement 70 (2021). doi:10.1109/TIM.2021.3116315

[28]Z. Li, W. Xie, L. Zhang, S. Lu, L. Xie, H. Su, W. Du, W. Hou, Toward efficient safety helmet detection based on yolov5 with hierarchical positive sample selection and box density filtering, IEEE Transactions on Instrumentation and Measurement 71 (2022). doi:10.1109/TIM.2022.3169564.